An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data

Puggini, L., & McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. *Engineering Applications of Artificial Intelligence*, *67*, 126-135. https://doi.org/10.1016/j.engappai.2017.09.021

Published in:
Engineering Applications of Artificial Intelligence

Document Version:
Publisher's PDF, also known as Version of record

An Enhanced Variable Selection and Isolation Forest Based Methodology for Anomaly Detection with OES Data

Luca Puggini[a], Seán McLoone[b]

[a]*National University of Ireland Maynooth, Maynooth, Ireland*
(e-mail: lpuggini@eeng.nuim.ie)
[b]*Queen's University Belfast, Belfast, England*
(e-mail: s.mcloone@qub.ac.uk)

Abstract

The development of efficient and interpretable anomaly detection systems is fundamental to keeping production costs low, and is an active area of research in semiconductor manufacturing, particularly in the context of using Optical Emission Spectroscopy (OES) data. The high dimension and correlated nature of OES data can limit the performance achievable with anomaly detection systems. In this paper we present a dimensionality reducing variable selection and isolation forest based anomaly detection and diagnosis methodology that addresses these issues. In particular, it takes account of isolated variables that can be overlooked when using conventional approaches such as PCA, and provides greater interpretability than afforded by PCA. The proposed methodology is illustrated with the aid of simulated and industrial plasma etch case studies.

Keywords: Semiconductors, Fault Detection, Dimensionality Reduction, OES Spectrum, Isolation Forest, Forward Selection Components Analysis

1. Introduction

Semiconductor manufacturing is one of the largest industries in the world, employing almost 250,000 people in the USA alone (Yinug, 2015). It posted sales globally totaling 335.2 billion dollars in 2015 (SIA, 2016). It is a highly competitive sector with manufacturers continually delivering new devices that are smaller, faster and/or more energy efficient than previous generations. Keeping pace with these developments, which have largely followed

Moore's law (Schaller, 1997), has resulted in the development of complex industrial processes, with product manufacture typically consisting of several hundred processing steps. Among these, plasma etching processes have been identified as critical to the production of semiconductor devices (Coburn and Winters, 1979). Figure 1 shows the main characteristics of a plasma etch process. Gases are pumped into a chamber where they are excited by microwaves to generate a plasma. The plasma then interacts with the exposed surface of the wafer both chemically and mechanically to etch away the wafer surface in a controlled fashion (Abe et al., 2008).

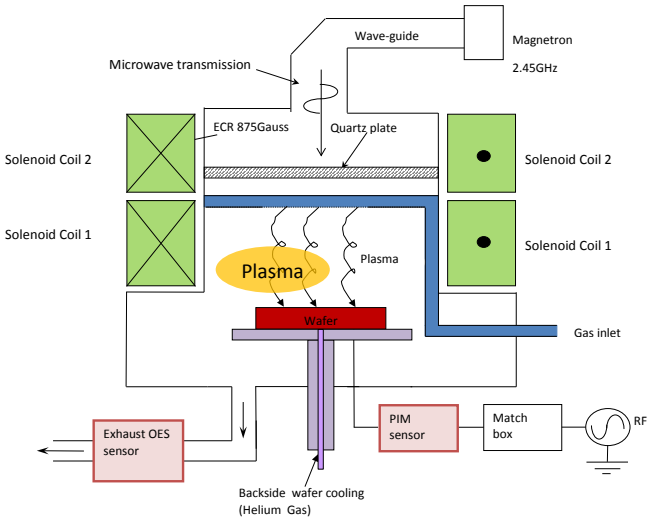

Figure 1: A plasma etching chamber.

Non-intrusive plasma monitoring can be achieved either by monitoring its electrical characteristics using a plasma impedance monitor (PIM) or by monitoring its optical output using optical emission spectroscopy (OES). OES monitoring is particularly attractive as it provides real-time information on the plasma chemical composition. This arises due to the unique wavelength signature that exists for each chemical species in the plasma. Data recorded from OES spectrometers consists of measurements of the intensity of the light emitted from the plasma as a function of a discrete set of wavelengths (channels) and time. As such, optical emission spectrometers are increasingly being deployed on plasma etch chambers to monitor plasmas, either directly through an optical window in the chamber, or indirectly through analysis of the exhaust gases from the chamber (as depicted in Figure 1). Figure 2

shows a typical spectrum generated by a plasma etch process. Several studies have shown OES to be an effective wafer processing monitoring signal (e.g. Chen et al. (1996), Puggini et al. (2014)). It has also been employed for applications such as fault detection (Yue et al., 2000) and etch rate prediction (Puggini and McLoone, 2015), (Zeng and Spanos, 2009).

Anomaly detection, in particular, is an active area of research in semiconductor manufacturing as the ability to detect faults early, and recognize anomalous behaviour in processes is key to improving product quality, overall process yield and throughput (He and Wang, 2007). Recent examples include Puggini et al. (2016) and Mahadevan and Shah (2009) who perform anomaly detection in OES time series data using unsupervised random forest and one class support vector machines (OC-SVM), respectively, and Ren and Lv (2014), He and Wang (2007) and Verdier and Ferreira (2011) who employ clustering based methodologies to separate normal and anomaly samples.

Anomaly detection with OES data is a challenging problem, due to its high dimension and highly correlated variables (Prakash et al., 2012). Both of these characteristics pose problems for anomaly detection algorithms. Most anomaly detection algorithms are based on a distance measure and it is known that such measures are unreliable in high dimensional spaces due to the so-called curse of dimensionality (Kriegel et al., 2008). As will be illustrated later in the paper, high levels of correlation among variables can degrade the performance of anomaly detection algorithms, as a small shift outside of the normal values in a group of correlated variables may generate a more anomalous result than a large shift in an isolated variable.

In this paper we propose a methodology for unsupervised anomaly detection and diagnosis using historical OES data that addresses both the dimensionality and correlation challenges. This consists of a dimensionality reduction pre-processing step, anomaly detection using the Isolation Forest algorithm (Liu et al., 2008), and a novel anomaly diagnosis procedure based on interrogation of the Isolation Forest (IF) model. In particular, building on our preliminary work in Puggini and McLoone (2016), we propose variable selection based dimensionality reduction techniques as a means of enhancing the interpretability of the IF model and improve its performance in the context of anomaly detection. In Puggini and McLoone (2016) we proposed a Forward Selection Independent Variables (FSIV) algorithm as an unsupervised variable selection technique specifically designed for anomaly detection. Here, we extend this work with more comprehensive industrial case studies, the introduction of a new variant of the algorithm that performs variable se-

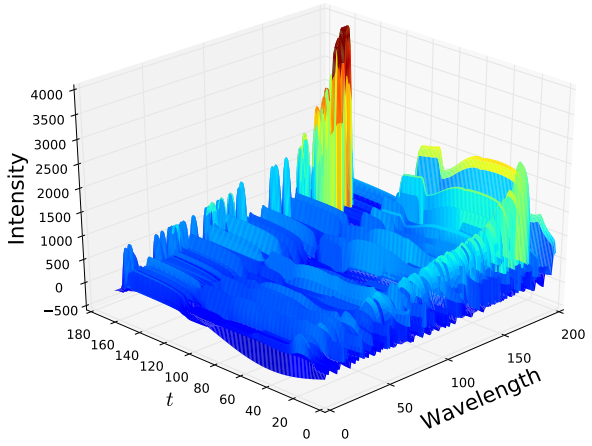

Figure 2: A typical of OES spectrum from the case study presented in Section 5

lection based on minimizing the maximum reconstruction error (and referred to as FSMM), and the development of a novel Isolation Forest based anomaly detection and diagnosis procedure.

The remainder of the paper is organized as follows. Section 2 introduces FSIV and FSMM and also briefly describes the underpinning Forward Selection Component Analysis (FSCA) (Puggini and Mcloone, 2017) algorithm. Section 3 provides an overview of IF and introduces the novel IF based fault diagnosis procedure. It also discusses the limitations of IF based anomaly detection with regard to correlated variables. A simulated example is then presented in Section 4 to illustrate this point, and to highlight the differences between FSIV, FSCA, and FSMM in the context of anomaly detection. For comparison purposes results are also presented for Principal Component analysis (PCA) (Jolliffe, 2002). The algorithms are then evaluated on two plasma etch process industrial cases studies in Sections 5 and 6. In the first case study OES time series data is available for a plasma etch process where a chamber seasoning effect is known to cause significant performance issues. In the second case study OES time series summary statistics are available for each wavelength for a plasma etch chamber that exhibits faulty behaviour, as evidenced by ground truth etch rate metrology data that is also available. Through these case studies the performance of the various unsupervised variable selection methods with IF based anomaly detection are compared and the application of the anomaly diagnosis procedure demonstrated. Finally,

conclusions are presented in Section 7.

2. Dimensionality Reduction in Anomaly Detection

Dimensionality reduction techniques such as PCA and FSCA seek to obtain lower dimensional approximations of datasets that retain the majority of the information in the original high dimensional datasets, usually defined in terms of the percentage of explained variance. PCA subsequently selects linear combinations of variables and FSCA sequentially selecting individual variables in order to maximize the explained variance at each iteration. While they are generally very useful for generating compact representations of highly correlated datasets, the reduced representations are not guaranteed to retain sufficient information to detect isolated anomalies (Puggini and McLoone, 2016). In particular, in datasets with several large clusters of correlated variables, the contributions of isolated uncorrelated variables to explained variance may be insignificant, with the result that such variables may not be included in the reduced data representation. It is then not possible to detect an anomaly if it is only reflected in such isolated variables.

Mitra et al. (2002) and Flynn and McLoone (2011) proposed recursive clustering algorithms that perform unsupervised feature selection that can retain isolated variables in the data. In the former, for each variable the set of its k-nearest variables is computed according to a similarity function. The variable which is closest to its k^{th} neighbour is retained while its k neighbours are discarded. The process ends when all the k-neighbours of all the variables are closer than a certain threshold to their centroid. In the latter, centroids for new clusters are chosen based on how different they are from the data in existing clusters, and individual clusters are formed on the basis of exceeding a similarity threshold. Then when clustering is complete the reduced dataset representation is defined as the centroids of the clusters. However, both these approaches select features based on a function $s(x, y)$ that measures the similarity between two variables. Higher order variable interactions are not considered. To address this, instead of discarding variables that are similar to those already selected FSIV analysis was proposed in Puggini and McLoone (2016) as a tool for efficient unsupervised features selection in anomaly detection. This begins with several iterations of the FSCA algorithm, before switching to a process of sequentially selecting the variables that are the least well represented by the selected variables. A natural extension of this concept is to sequentially select the variables that minimize

the maximum variable reconstruction error at each iteration. Hereafter, this algorithm variant will be referred to as FSMM.

Figures 3, 4 and 5 show the steps required to select K variables using FSCA, FSIV and FSMM, respectively. In each case it is assumed that the data has been scaled to have zero mean and unit variance. FSCA is the base algorithm. It selects K variables with the aim of maximizing the percentage of explained variance. Both FSIV and FSMM are initialized by employing FSCA to select an initial set of k_1 variables. They then select a further k_2 variables based on the maximum reconstruction error over all the variables. In FSIV at each iteration this is simply taken as the variable with the largest error following reconstruction by the current set of selected variables, while in FSMM it is the variable which leads to the smallest maximum reconstruction error when combined with the currently selected variables. In both algorithms the initial FSCA step is included to ensure that the variables representing the largest variation in the data are included. Then, additional variables are added in order to include significant isolated variations that are not captured by the first k_1 variables.

The distinguishing feature of FSIV and FSMM is that a variable is added to the model if it cannot be adequately reconstructed by a linear combination of those already selected. This makes these algorithms more efficient than methods based on similarity between variables. This follows, for example, from the fact that weakly correlated variables may be linearly dependent (Rodgers et al., 1984).

2.1. Performance Metrics

Given the original matrix $X \in \mathbb{R}^{n \times p}$ and a lower dimensional approximation $Z \in \mathbb{R}^{n \times k}$ the approximation of X is defined as:

$$\hat{X} = Z(Z^T Z)^{-1} Z^T X \in \mathbb{R}^{n \times p} \tag{10}$$

The quality of this approximation can be evaluated with a series of metrics. The percentage normalized mean squared error between X and its approximation can be expressed as:

$$E_{NMSE} = 100 \times \frac{\| X - \hat{X} \|_2^2}{\| X \|_2^2} \tag{11}$$

FSCA Algorithm

$Z_K = \mathbf{FSCA}(X, K)$:

1. Start with the full data $X = (x_1, \ldots, x_p)$ and K the number of variables to select. Initialize $Z_0 = \emptyset$ and $k = 0$.
2. Define Z_{k+1}^v as the matrix Z_k with the addition of the variable x_v i.e. $Z_{k+1}^v = (Z_k, x_v)$
3. Define Z_{k+1} as:

$$argmin_v \parallel X - Z_{k+1}^v (Z_{k+1^v}^T Z_{k+1}^v)^{-1} {Z_{k+1}^v}^T X \parallel_2 \tag{1}$$

4. Update $k = k + 1$
5. If $k < K$ return to step 3. Otherwise output Z_K, the set of selected variables.

Figure 3: Pseudocode for the Forward Selection Component Analysis (FSCA) algorithm

with the corresponding percentage of explained variance (EV) given by:

$$EV = 100 - E_{NMSE} \tag{12}$$

E_{NMSE} and EV measure the average reconstruction performance over all the variables in X.

In contrast, the Maximal Reconstruction Error (E_{MRE}) metric used in FSIV and FSSM only considers the variable with the largest reconstruction error and is computed as

$$E_{MRE} = 100 \times p \times max_v \frac{\parallel x_v - \hat{x}_v \parallel_2^2}{\parallel X \parallel_2^2} \tag{13}$$

The factor p is applied to provide equivalent normalisation to E_{NMSE}. From this definition it follows that

$$E_{MRE} \geq E_{NMSE} \tag{14}$$

with the equivalence only holding if all variables have the same reconstruction error.

FSIV Algorithm

$Z = \mathbf{FSIV}(X, k_1, k_2)$ where $k_1 + k_2 = K$:

1. Start with the full data $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$
2. Select k_1 variables $z_1, \ldots, z_{k_1}$ using the FSCA algorithm.
3. Define the matrix $Z = (z_1, \ldots, z_{k_1})$.
4. Compute the linear approximation of X

$$\hat{X} = Z(Z^T Z)^{-1} Z^T X \in \mathbb{R}^{n \times k} \tag{2}$$

where

$$\hat{X} = (\hat{x}_1, \ldots, \hat{x}_p) \tag{3}$$

5. For each variable x_i in X compute its approximation error

$$\epsilon_i = \| x_i - \hat{x}_i \|_2^2 \tag{4}$$

where $\hat{x}_i$ is the ith column of $\hat{X}$.

6. Select $x_{\hat{j}}$ the variable with the highest approximation error where:

$$\hat{j} = \underset{i}{\operatorname{argmax}}\, \epsilon_i \tag{5}$$

7. Add $x_{\hat{j}}$ to the Z matrix.
8. Stop if $K = k_1 + k_2$ variables have been selected, otherwise repeat from step 4

Figure 4: Pseudocode for the Forward Selection Independent Variable (FSIV) algorithm

FSMM Algorithm

$Z = \mathbf{FSMM}(X, k_1, k_2)$, where $k_1 + k_2 = K$.

1. Select k_1 variables $z_1, \ldots, z_{k_1}$ using the FSCA algorithm.
2. Define the matrix $Z = (z_1, \ldots, z_{k_1})$.
3. Define Z^v as the matrix Z with the addition of the variable x_v i.e. $Z^v = (Z, x_v)$
4. Obtain $\hat{X}^v$ as:
$$\hat{X}^v = Z^v(Z^{vT}Z^v)^{-1}Z^{vT}X \tag{6}$$
5. Compute the reconstruction error $\hat{\epsilon}_k^v$ for each variable and the maximal reconstruction error $\hat{\epsilon}^v$ as:
$$\hat{\epsilon}_k^v = \parallel \hat{x}_k^v - x_k \parallel_2 \quad for\ k = 1, \ldots, p \tag{7}$$
$$\hat{\epsilon}^v = max_k \hat{\epsilon}_k^v \tag{8}$$
6. Save the variable $\hat{v}$ that leads to the smallest maximal reconstruction error.
$$\hat{v} = argmin_v \hat{\epsilon}^v \tag{9}$$
7. Update Z as the matrix Z with the addition of the variable $x_{\hat{v}}$ i.e. $Z = (Z, x_{\hat{v}})$
8. If a total of $K = k_1 + k_2$ variables have been selected output Z, the set of selected variables. Alternatively return to step 3.

Figure 5: Pseudocode for the Forward Selection Minimizing the Maximum Reconstruction Error (FSMM) algorithm

3. Fault Detection and Diagnosis with Isolation Forest

An isolation forest is an ensemble of isolation trees, similar to the more popular decision trees and random forest (Murthy, 1998) and (Breiman, 2001). An isolation tree is constructed starting with a matrix X as described in Figure 6. An Isolation Forest is then defined by numerous Isolation Trees

$$IF = \{t_1, \ldots, t_T\} \tag{15}$$

For each tree t it is possible to compute the number of iterations $h_t(x)$ required to isolate a sample x. The average number of steps required to isolate a sample x in a forest is then

$$h(x) = \frac{1}{T} \sum_{t \in IF} h_t(x) \tag{16}$$

The idea is that only a few steps are required to isolate an anomaly. The number of steps required to isolate an observation x is influenced by the number of samples n in the data. To account for this a normalized anomaly score $s(x, n)$ is defined as:

$$s(x, n) = 2^{-\frac{h(x)}{c(n)}} \tag{17}$$

where $c(n)$ is:

$$c(n) = \begin{cases} 2H(n-1) - 2(n-1)/n & \text{if } n > 2 \\ 1 & \text{if } n = 2 \\ 0 & otherwise \end{cases} \tag{18}$$

and $H(i)$ is the harmonic number estimated as:

$$H(i) \approx ln(i) + 0.5772156649. \tag{19}$$

It can be proven that $c(n)$ is the average number of steps required to isolate a sample from the other n samples (Liu et al., 2008). In this sense it provides a normalization factor that makes the s value independent of the number of samples (n).

Isolation Tree Algorithm

Require: Input matrix $X \in \mathbb{R}^{n \times p}$.

1: $t = \emptyset$ (the empty tree)
2: **if** $nrow(X) == 1$ **then return** t
3: **end if**
4: Randomly select x_i a feature of X
5: Randomly select a split point $p \in (\min(x_i), \max(x_i))$
6: Add to t the node $N_{x_i,p}$
7: Define X_l and X_r as the matrix composed of the samples of X where the variable x_i is respectively larger and smaller than p.
8: Repeat the algorithm with $X = X_l$. Link the obtained tree as the left child of t.
9: Repeat the algorithm with $X = X_r$. Link the obtained tree as the right child of t.

Figure 6: Psuedocode for the Isolation Tree algorithm

3.1. Anomaly Threshold

Once an anomaly score has been assigned to each sample, it is necessary to define a threshold beyond which an observation is considered anomalous and flagged. We have observed that the anomaly score tends to follow an F distribution (see for example Figures 8 and 12 for the case studies presented later). Therefore, in order to detect anomalies, once the anomaly score is computed for the set of training samples, an F distribution is used to approximate the data. Anomalies are then considered as those observations with an anomaly score that is greater than the right limit of the confidence interval at a given percentile.

3.2. Diagnosis Procedure:

While tree based methods are generally considered black box algorithms, the following procedure may be used to detect the set of variables that are causing the anomaly. Given the anomaly $a \in \mathbb{R}^p$ we can define $h(a)$ as the average number of splits that are required to isolate a. Consider the two subsets of variables $X_i = \{x_{i_1}, \ldots, x_{i_k}\}$ and $X_j = \left\{x_{j_1}, \ldots, x_{j_{(p-k)}}\right\}$ such that $X = (X_i, X_j)$. X_i is the set of variables from which subsets of variables can be extracted which can isolate a with a few splits. X_j instead contains

the set of variables from which it is more difficult to extract a subset that can isolate a in a few splits. Since

$$h(a) = \frac{1}{T} \sum_{t \in IF} h_t(a) \tag{20}$$

it is clear that trees which have as an initial split variables that are mainly from set X_j require more steps to isolate a than ones that have most of their initial split variables from the set X_i. Using this observation it is possible to detect which variables are causing the anomaly by analysing the most frequently occurring split variables that appear in the initial positions of the trees that isolate the anomaly with a low number of splits.

3.3. Bias Toward Correlated Variables

IF is an effective and computationally efficient anomaly detection algorithm that scales well to high dimensional datasets (Liu et al., 2008). However, one weakness is that it is biased with respect to groups of correlated variables. Given two anomalies a_1 and a_2 of which the former is an anomaly respect to a group of correlated variables while the latter is an anomaly respect to a single variable the IF algorithm will tend to assign an higher anomaly score to a_1. The reason for this follows from the algorithm that is used to build the Isolation Trees as at each split it is more likely to have a variable that can isolate a_1 rather than a_2. As a consequence preprocessing the data with algorithms that reduce the number of correlated variables such as FSCA or FSIV will result in an improvement in performance of the IF algorithm.

4. Simulated Example

The following simulated example is used to illustrate the difference between the discussed dimensionality reduction algorithms (FSMM, FSIV, FSCA and PCA) and the bias of IF towards groups of correlated variables.

4.1. The Data

Consider the simulated data $X = (x_1, \ldots, x_7) \in \mathbb{R}^{n \times 7}$ with $n = 1000$ samples and defined by three groups of variables $X_1 = \{x_1, x_2, x_3\}$, $X_2 = \{x_4, x_5, x_6\}$ and $X_3 = \{x_7\}$. Each variable has correlation 0.9 with the others

in the same group and between the variables in X_1 and X_2 there is a correlation of 0.4. The variable in X_3 is instead isolated and has only correlation 0.1 with all other variables. Given $\Sigma = \{\Sigma_{i,j}\} \in \mathbb{R}^{7\times 7}$ defined as:

$$\Sigma_{i,j} = \begin{cases} 1 & if\ i = j \\ 0.9 & if\ i,j \in \{1,2,3\}\ or\ i,j \in \{4,5,6\} \\ 0.4 & if\ i \in \{1,2,3\}\ and\ j \in \{4,5,6\} \\ 0.4 & if\ j \in \{1,2,3\}\ and\ i \in \{4,5,6\} \\ 0.1 & if\ i = 7\ or\ j = 7 \end{cases}$$

the matrix $X = (x_1, \ldots, x_7) \sim N(0, \Sigma)$ is given by

$$X = LU \tag{21}$$

where $U \in \mathbb{R}^{1000\times 7}$ is a matrix whose elements are independent random samples of a $N(0,1)$ distribution and L is obtained from the Cholesky decomposition Wilkinson (1965) of Σ

$$\Sigma = LL^T \tag{22}$$

An anomaly is then introduced by replacing one of the samples in x_7 with the value 10.

4.2. Dimensionality Reduction

FSCA, FSIV and FSMM are employed to perform dimensionality reduction. In each case only two variables are selected. In FSIV and FSMM the parameters are $k_1 = k_2 = 1$. A two component PCA representation of the data is also computed for comparison purposes. The dimensional representations of the data obtained with the various methods are reported in Figure 7. Here, the blue dots represent the normal data points and the red star signifies the single anomalous data point. FSIV and FSMM select the same variables and as a consequence they have identical results. From the figure it can be observed that only FSIV and FSMM are able to isolate the anomaly. In particular, FSCA tends to select one variable from X_1 and one from X_2 while FSIV and FSMM select a variable from X_1 and x_7. The PCA components instead are obtained as a weighted linear combination of all the variables. However, the weighting associated with x_7 is insufficient to materially affect the behaviour of the components, with the result that the anomaly is not distinguishable from the normal samples.

4.3. Anomaly Detection

Once a lower dimension approximation of the data is obtained, it is processed using IF and an anomaly score generated for each sample. Figure 8 shows the distribution of the obtained anomaly scores on the data reduced with the FSIV/FSMM algorithms. The figure also shows the probability density function of an F distribution which has been fitted to the data. This is used to define confidence intervals for the anomaly score. Figure 9 shows the anomaly score and the control limits obtained with the F distribution when all the variables are used and when the data was reduced with FSIV, FSMM, FSCA and PCA. The last sample is the one containing the anomaly and it is interesting to observe that this has an anomaly score outside of the 99.9% confidence interval only if the data is first reduced with the FSIV or FSMM algorithms. The abnormal sample has a very low anomaly score if the data is reduced with FSCA or PCA and is indistinguishable from the normal data. This is easily follows from Figure 7 as the anomaly can not be distinguished by the normal samples in the obtained lower dimensional representation of the data. When all the variables are used the abnormal sample does exceed the 99% confidence threshold, but it has a lower anomaly score than three normal behaving samples (false alarms). This is due to the the IF bias towards groups of correlated variables as explained in Section 3.3.

5. Industrial case study 1: OES Time Series

To demonstrate the effectiveness of the proposed dimensionality reduction and anomaly detection method, the technique is applied to a sample OES dataset collected over several months from an industrial plasma etch chamber which is subject to a significant chamber seasoning affect. In the production line wafers are grouped in lots of 25 wafers, with wafers in a lot arranged in slots on a cassette. Wafers in a given lot are processed sequentially (according to slot number). Lots are also processed sequentially through the etch chamber, interspersed with cleaning and maintenance operations. Cleaning cycles are typically done between each lot to remove the by-products of plasma etching that build up on the chamber walls, and are detrimental to etching performance. This leads to a chamber seasoning effect during the first few wafers processed following each cleaning cycle, and consequently slot dependent differences in processed wafers. In particular, in this case study the plasma etching of the wafers in slot 1 and 2 are deemed to be anomalous with respect to the wafers in the other slot positions.

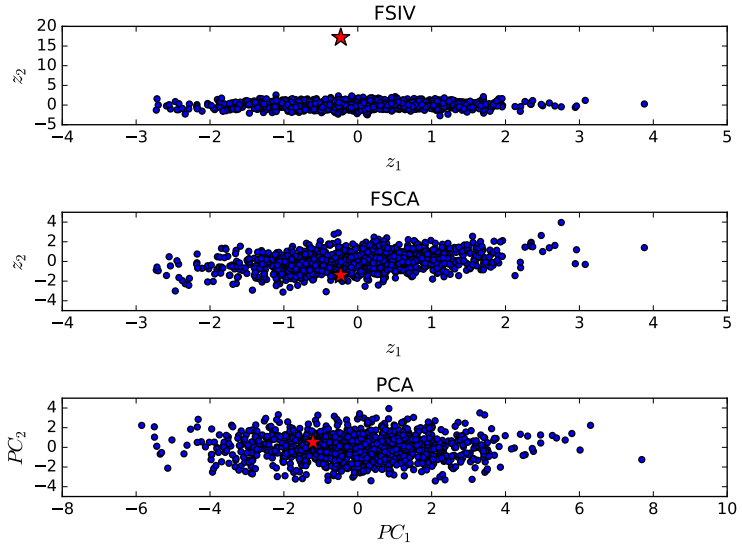

Figure 7: The projection of the data for the simulated example on the first two variables selected by FSCA and FSIV, and the first two principal components obtained with PCA. The blue dots are the normal data and the red star is the data point with the anomaly.

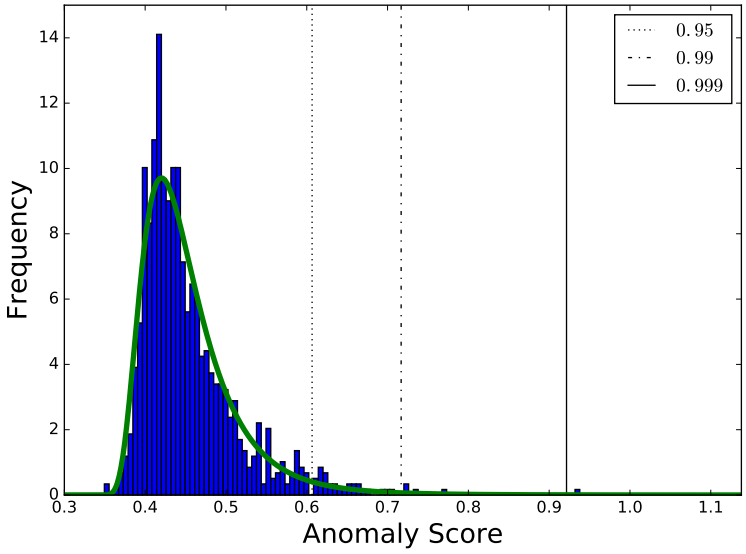

Figure 8: The distribution of the anomaly score when the data is preprocessed with FSIV and the probability density function of an F distribution and the right limit of confidence intervals 0.95, 0.99 and 0.999 confidence intervals.

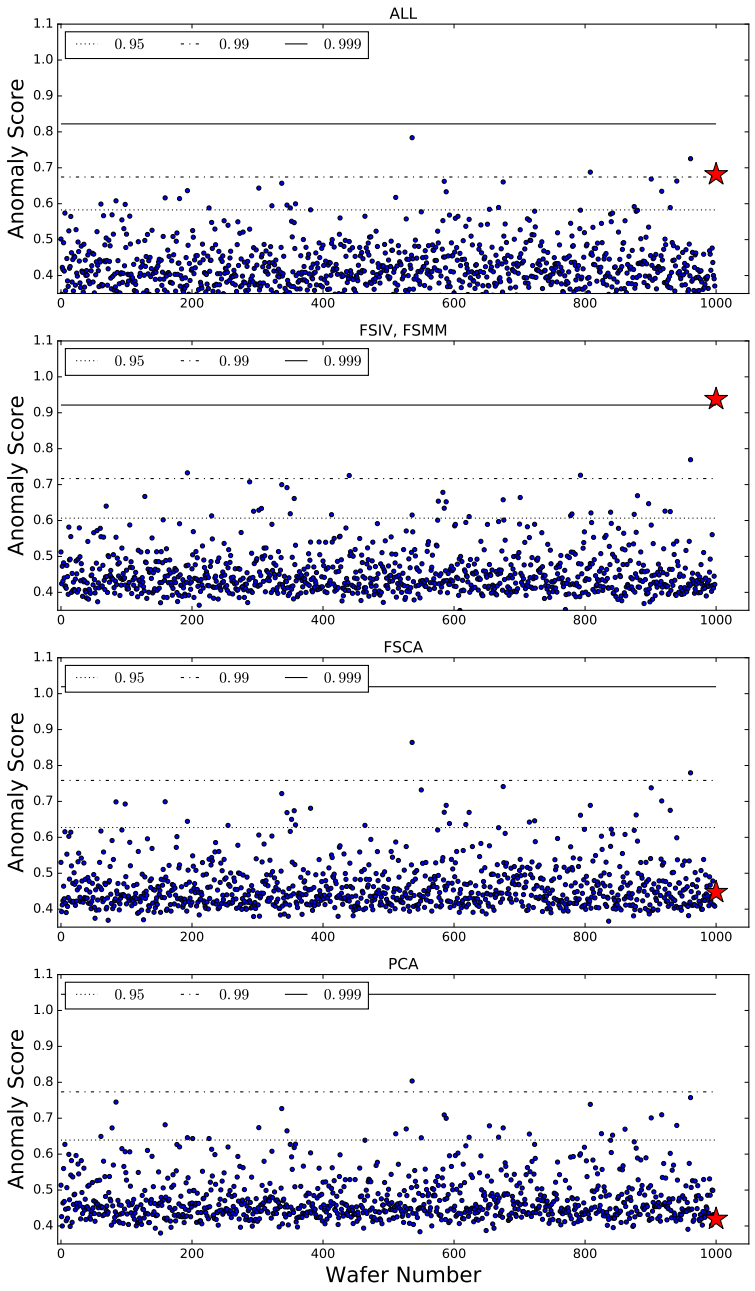

Figure 9: The Anomaly Score obtained with Isolation Forest on the simulated example when all the variables are used (plot 1), and when the dimensionality of the data is reduced with FSIV/FSMM, FSCA and PCA (plots 2, 3 and 4, respectively). The red star denotes the abnormal data point and the blue dots are the normal data. Only FSIV/FSMM provide effective discrimination of the anomaly.

Case Study 1: OES Time Series	
Dataset	1000 wafers x 500 wavelengths x 20 time samples
Normal	920 wafers (wafers 3-25 in each lot processed)
Abnormal	80 wafers (wafers 1 and 2 in each lot processed) with inferior etch performance due to a transient chamber seasoning effect following the cleaning cycle performed before each lot of wafers is processed

Case Study 2: OES Summary Statistics	
Dataset	1500 wafers x 500 wavelengths x 6 summary statistics
Normal	Wafers with measured etch rate (ER) falling between 65 and 75 (normalised units)
Abnormal	a) 5 wafers in the vicinity of wafer number 200 with ER $<<$ 65 due to a wafer cooling process fault b) 91 wafers with ER $>$ 75 between wafer number 950 and 1380 due to a process shift following a maintenance intervention

Table 1: A overview of the characteristics of the industrial datasets used as case studies in Sections 5 and 6

The available dataset consists of Plasma Etch Optical Emission Spectroscopy (OES) time series recordings from the chamber exhaust (as depicted in Fig. 1). Noting that the OES data is naturally parameterized in terms of the wafer number, the processing time instant and the measured wavelengths (Yue et al., 2000), the intensity of the i^{th} wavelength of the k-th wafer at time t is denoted as $x_i^{w_k}(t)$. OES spectra are available for $N = 1000$ wafers, with each one consisting of $\tau = 20$ samples of $p = 500$ wavelengths. The dataset characteristics are summarised in Table 1. The OES spectrum for a single wafer w_k can be mathematically represented as a matrix $X_k \in \mathbb{R}^{\tau \times p}$.

$$X_k = \left\{ x_i^{w_k}(t_{(k-1)\tau+j}) \right\}_{j=1,\ldots,\tau,\ i=1,\ldots,p} \in \mathbb{R}^{\tau \times p} \tag{23}$$

and the full data is represented by a set S containing the measurements for each wafer:

$$S = \left\{ X_k \in \mathbb{R}^{t_\tau \times p} :\ k = 1, \ldots, N \right\} \ . \tag{24}$$

For analysis purposes it is often desirable to merge the data from all wafers in a dataset into a single two dimensional matrix representation. Two possible aggregations are considered and are denoted as $\Lambda \in \mathbb{R}^{\tau N \times p}$ and $W \in \mathbb{R}^{N \times p\tau}$.

5.1. The Λ matrix

The data can be aggregated in a $\Lambda \in \mathbb{R}^{\tau N \times p}$ matrix. In Λ each column corresponds to the light intensity of a specific wavelength (spectrometer channel) and each row is its uniformly sampled evolution over time. Λ can be obtained by vertically stacking the matrices in S.

$$\Lambda = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_N \end{pmatrix} \in \mathbb{R}^{N\tau \times p} \tag{25}$$

The full OES data is then represented as the concatenation of a set of matrices where each one contains the spectrum for a given wafer. This format of the data is used with the dimensionality reduction techniques to determining the most appropriate subset of wavelengths to represent the complete dataset.

5.2. W Matrix

Alternatively the data can be aggregated in order to have each wafer as an observation. Here the matrices representing each wafer are converted into

row vectors (by concatenating each column vector in the matrix and then transposing it) and stacking them. Specifically, the k_{th} element of S, matrix X_k, is reshaped as

$$\tilde{X}_k = (x_1^{w_k}(t_1), \ldots, x_p^{w_k}(t_\tau)) \in \mathbb{R}^{1\times\tau p} \tag{26}$$

and the full dataset is represented by combining all the reshaped matrices in S as:

$$W = \begin{pmatrix} \tilde{X}_1 \\ \cdots \\ \tilde{X}_N \end{pmatrix} \in \mathbb{R}^{N\times\tau p} \tag{27}$$

This data format is particularly useful for comparing wafers and performing anomaly detection as each row corresponds to all the data observed for a given wafer. However, the dimension of the matrix is generally very large. For example, it has $p\tau = 10000$ columns for the current case study. The dimension of the data can be drastically reduced by selecting only a subset of the wavelengths. If K wavelengths are selected based on analysis of the Λ format of the data, the number of columns in the W matrix is reduced to $K\tau << p\tau$ columns.

5.3. Dimensionality Reduction and Anomaly Detection Performance

K variables are selected with PCA, FSCA, FSIV and FSMM from the data represented in the $\Lambda \in \mathbb{R}^{N\tau\times p}$ format. The obtained lower dimensional representation $Z \in \mathbb{R}^{N\tau\times K}$ is then transformed into the W format and IF is used to assign an anomaly score to each wafer. By considering the wafers in slot 1 and 2 as anomalous due the post cleaning cycle chamber seasoning transient the ability to distinguish them from the wafers processed in the remaining slots provides a useful ground truth for evaluating the performance of the proposed anomaly detection system. The results obtained for this scenario are given in Table 2 for different values of K (the size of the low dimension approximation to the 500 dimensional OES dataset), and in the case of FSIV and FSMM for different k_1, k_2 combinations. In addition, Figure 10 shows box plots of the anomaly score distribution for the wafers as a function of slot position for the case where $K = 10$, $k_1 = 3$ and $k_2 = 7$.

The reconstruction error on the Λ matrix reported in Table 2 are measured in terms of EV, E_{NMSE} and E_{MRE}, as defined in Section 2.1, while the anomaly detection performance is expressed in terms or the standard binary classifier AUC (Area Under the receiver operating Curve) metric (Bradley,

1997). As expected for a given number of components K, PCA yields the maximum explained variance (EV), or equivalently the lowest E_{NMSE}. However, it does not always yield the smallest E_{MRE}. When $K = 2$ and $K = 10$, for example, FSMM achieves better E_{MRE} performance. More importantly, the variable selection based dimensionality reduction methods, and in particular FSIV and FSMM, yield IF models that consistently outperform the PCA model in terms of anomaly detection capability. The best AUC score is obtained with FSIV or FSMM for all the values of K while PCA always yields the worst score. FSCA falls between these extremes in terms of AUC performance achieving AUC scores than are better than PCA but worse than FSIV or FSMM. As FSCA selects variables on the basis of maximizing the explained variance it achieves EV and E_{NMSE} performance approaching that of PCA. While in general it is second only to PCA in this regard, for $K = 5$ FSIV(1,4) is marginally superior. This anomaly arises as FSCA employs a greedy search method for variable selection which is not guaranteed to find the global optimum solution. Comparing FSIV and FSMM the pattern is less consistent with FSMM outperforming FSIV for $K = 2$ and 5 and vice versa for $K = 10$ and 15. However, it should be noted that differences between methods becomes less significant with increasing K.

Figure 10 also clearly shows that the IF based anomaly score is able to capture the chamber seasoning effect, with the post cleaning transient clearly evident, as well as the degradation in performance at high slot numbers due to plasma etch by-product buildup. It is also evident that PCA yields the narrowest spread of anomaly scores as a function of slot number with FSMM and FSIV yielding the largest spread. This corroborates the AUC pattern observed in Table 2.

6. Industrial Case Study 2: OES Summary Statistics

The second case study is an OES dataset obtained for a plasma etching chamber which experiences a process fault and later a maintenance related process shift during production. This dataset consists of OES summary statistics for 1500 wafers. The dataset characteristics are summarized in Table 1. To reduce data volume OES summary statistics were recorded for the time evolution of each wavelength for the processing step, rather than raw time series values. The summary statics used were mean, variance, skewness, kurtosis, minimum and maximum values. The etch rate for each wafer was measured in a post-processing metrology step and is plotted in Figure

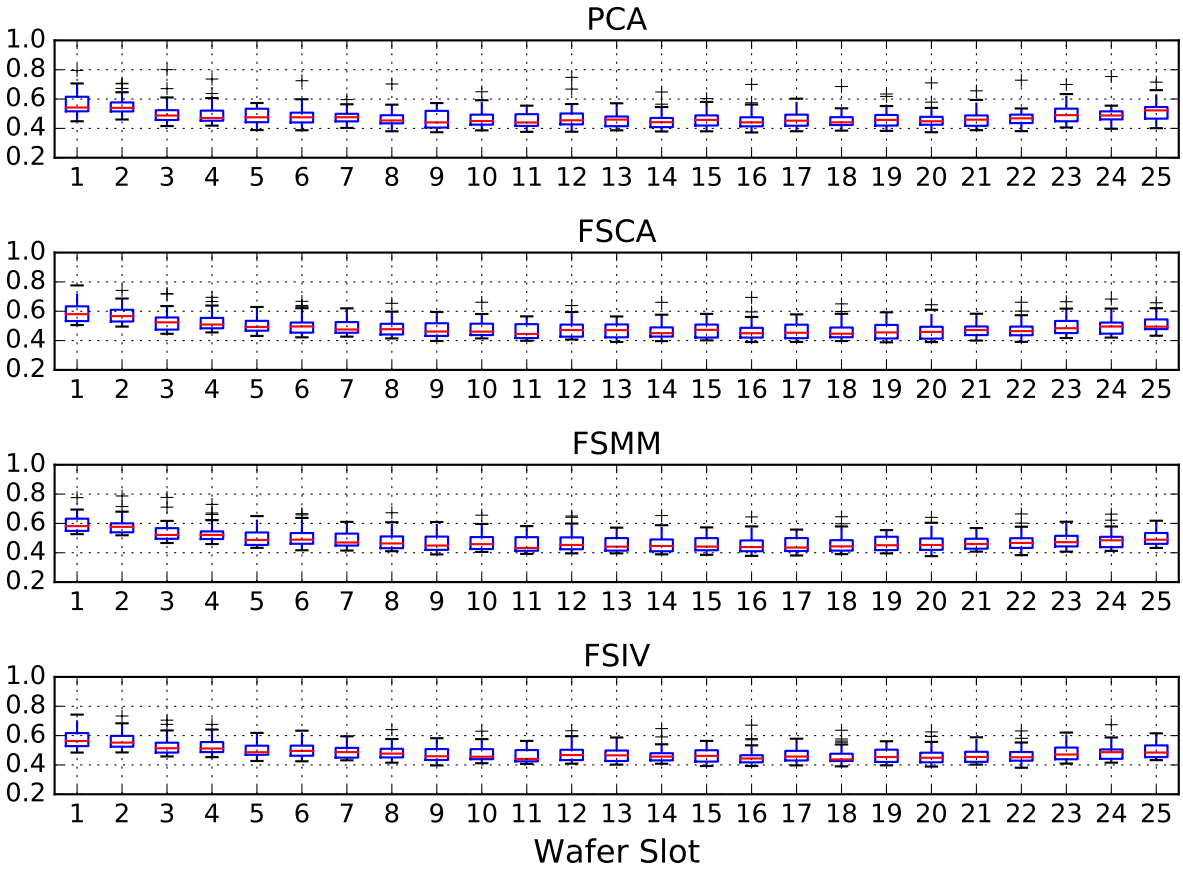

Figure 10: The anomaly score assigned to each wafer according to its Wafer Slot.

	EV	E_{NMSE}	E_{MRE}	AUC
		$K = 2$		
PCA	94.21	5.79	90.35	0.87
FSCA	93.62	6.38	91.61	0.90
FSIV(1,1)	93.11	6.89	96.47	0.87
FSMM(1,1)	93.23	6.77	83.20	**0.93**
		$K = 5$		
PCA	98.94	1.06	24.48	0.86
FSCA	98.34	1.66	39.39	0.88
FSIV(1,4)	98.38	1.62	25.32	0.90
FSMM(1,4)	97.38	2.62	42.81	**0.93**
FSIV(2,3)	98.11	1.89	24.72	0.89
FSMM(2,3)	97.53	2.47	40.12	0.90
FSIV(3,2)	98.22	1.78	25.91	0.90
FSMM(3,2)	98.22	1.78	37.10	0.91
FSIV(4,1)	98.20	1.80	26.06	0.89
FSMM(4,1)	98.20	1.80	26.06	0.89
		$K = 10$		
PCA	99.28	0.72	13.94	0.84
FSCA	98.98	1.02	23.96	0.88
FSIV(1,9)	98.78	1.22	10.51	0.91
FSMM(1,9)	98.66	1.34	19.33	0.89
FSIV(3,7)	98.67	1.33	10.83	**0.92**
FSMM(3,7)	98.77	1.23	13.73	0.91
FSIV(5,5)	98.77	1.23	11.38	0.90
FSMM(5,5)	98.91	1.09	11.46	0.89
FSIV(7,3)	98.91	1.09	11.46	0.90
FSMM(7,3)	98.91	1.09	11.46	0.90
FSIV(9,1)	98.96	1.04	19.89	0.90
FSMM(9,1)	98.96	1.04	19.89	0.90
		$K = 15$		
PCA	99.42	0.58	7.48	0.87
FSCA	99.23	0.77	10.18	0.91
FSIV(1,14)	98.97	1.03	8.73	**0.92**
FSMM(1,14)	99.07	0.93	8.84	0.91
FSIV(5,10)	98.99	1.01	8.42	0.91
FSMM(5,10)	99.12	0.88	9.51	0.90
FSIV(7,8)	99.07	0.93	8.60	0.89
FSMM(7,8)	99.12	0.88	9.51	0.90
FSIV(10,5)	99.17	0.83	9.15	0.91
FSMM(10,5)	99.19	0.81	9.60	0.91
FSIV(14,1)	99.23	0.77	10.18	0.91
FSMM(14,1)	99.23	0.77	10.18	0.91

Table 2: EV, E_{NMSE}, E_{MRE} and the AUC score obtained with IF and the different dimensionality reduction methods for case study 1, as described in Section 5

11. Normal etch rate is defined to be in the range $[66, 75]$. From the etch rate data the process fault is clearly evident in the vicinity of wafer index 200 while the process shift is evident between wafer 950 and wafer 1380. It should be noted that etch rate data is not normally available for each production wafer as it is costly and time-consuming to measure and is often only sparsely sampled and available several hours or days after production. Hence, the challenge is to recognize the abnormal process behaviour using only the OES data which is collected for each wafer during production.

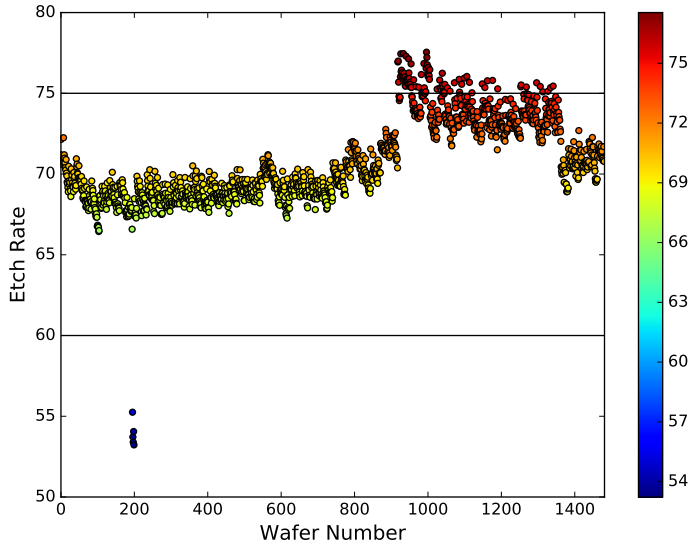

Figure 11: The etch rate measurements for case study 2. The horizontal lines indicate the normal operating range for etch rate values.

6.1. Dimensionality Reduction and Anomaly Detection

Following preprocessing to remove inactive and low signal-to-noise ratio wavelengths the resulting data matrix available for analysis had 500 variables, i.e. $X \in \mathbb{R}^{1500\times500}$. Each variable was scaled to zero mean and unit variance. The various dimensionality reduction techniques and IF based anomaly detection methodology were applied to this dataset and the results obtained are as reported in Table 3. The results are similar to those observed in case study 1 with PCA explaining the most variance but delivering the worst AUC performance for each value of K considered. In contrast to the first case study FSMM outperforms FSIV for all values of K and yields the best

AUC results overall. In addition, FSCA also outperforms FSIV and achieves equivalent performance to FSMM for $K = 2$ and $K = 5$.

By way of illustration Figure 12 shows that distribution of IF anomaly scores obtained for the 1500 wafers and associated best fit F distribution when $K = 10$, $k_1 = 3$, and $k_2 = 7$. The anomaly scores and the thresholds derived from the F distribution are plotted in Figure 13, with each sample colour coded to indicate the corresponding metrology etch rate. As can be seen the faulty wafers positioned near index 200 are clearly marked as anomalies by all algorithms. In contrast the process shift anomaly is less distinguishable from normal samples with all methods. PCA in particular results in significant overlap with the anomaly scores of the normal wafers at the start of the dataset, while FSMM yeilds marginally the best separation. The is reflected in an AUC score of 0.88 for PCA and 0.91 for FSMM in Table 3. The difficulty IF has isolating the process shift samples is due to their large number and their proximity to normal samples, i.e. they are not sufficiently anomalous.

To illustrate the application of the fault diagnosis procedure one of the faulty samples in the vicinity of index 200 is selected for investigation. Figure 14 shows the histogram of the number of splits required to isolate this sample from the rest of the data. As can be seen only a small number of splits are needed for the majority of trees in the IF with each dimensionality reduction technique. Figure 15 shows the data in the subspace defined by variables x_2 and x_8. These were variables used as split points in an isolation tree that was able to isolate the samples with 2 splits. As can be seen the faulty wafers are clearly distinguishable in this plot.

	EV	E_{NMSE}	E_{MRE}	AUC
		$K = 2$		
PCA	72.79	27.21	79.28	0.88
FSCA	70.46	29.54	89.01	**0.90**
FSIV(1,1)	63.79	36.21	99.55	0.86
FSMM(1,1)	69.18	30.82	78.84	**0.90**
		$K = 5$		
PCA	94.00	6.00	29.64	0.87
FSCA	91.40	8.60	44.66	**0.89**
FSIV(1,4)	90.76	9.24	46.74	0.88
FSMM(1,4)	84.54	15.46	42.41	0.88
FSIV(2,3)	88.46	11.54	41.22	0.86
FSMM(2,3)	89.87	10.13	41.19	0.88
FSIV(3,2)	89.10	10.90	46.34	0.87
FSMM(3,2)	89.89	10.11	41.08	**0.89**
FSIV(4,1)	90.29	9.71	44.98	0.88
FSMM(4,1)	90.41	9.59	43.93	**0.89**
		$K = 10$		
PCA	98.37	1.63	12.71	0.88
FSCA	97.23	2.77	22.12	0.90
FSIV(1,9)	95.98	4.02	13.69	0.86
FSMM(1,9)	94.61	5.39	19.60	0.89
FSIV(3,7)	96.33	3.67	12.50	0.87
FSMM(3,7)	95.14	4.86	18.28	**0.91**
FSIV(5,5)	95.94	4.06	17.55	0.86
FSMM(5,5)	95.70	4.30	20.00	0.90
FSIV(7,3)	96.64	3.36	19.28	0.89
FSMM(7,3)	96.51	3.49	17.43	0.89
FSIV(9,1)	97.21	2.79	22.89	0.90
FSMM(9,1)	97.22	2.78	22.00	0.89
		$K = 15$		
PCA	99.43	0.57	3.21	0.89
FSCA	98.77	1.23	7.10	0.89
FSIV(1,14)	98.72	1.28	5.73	0.90
FSMM(1,14)	97.63	2.37	10.84	0.89
FSIV(5,10)	98.69	1.31	5.38	0.90
FSMM(5,10)	97.65	2.35	12.16	**0.92**
FSIV(7,8)	98.70	1.30	7.09	0.89
FSMM(7,8)	98.41	1.59	7.18	0.91
FSIV(10,5)	98.67	1.33	7.34	0.90
FSMM(10,5)	98.59	1.41	6.88	**0.92**
FSIV(14,1)	98.73	1.27	7.58	0.88
FSMM(14,1)	98.69	1.31	6.68	0.90

Table 3: EV, E_{NMSE}, E_{MRE} and the AUC score obtained with IF and the different dimensionality reduction methods for case study 2, as described in Section 6

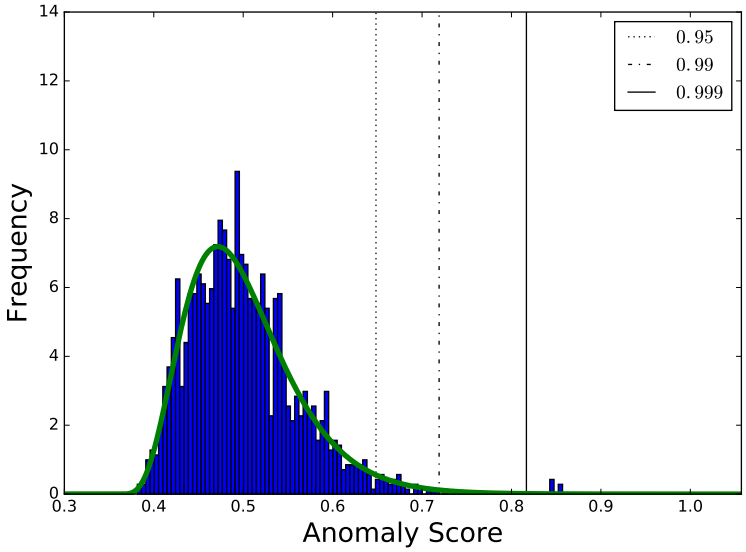

Figure 12: The distribution of the anomaly scores obtained with IF in case study 2 and the estimated F distribution (green) model. The three vertical lines represent the upper 95, 99 and 99.9% confidence intervals of the distribution.

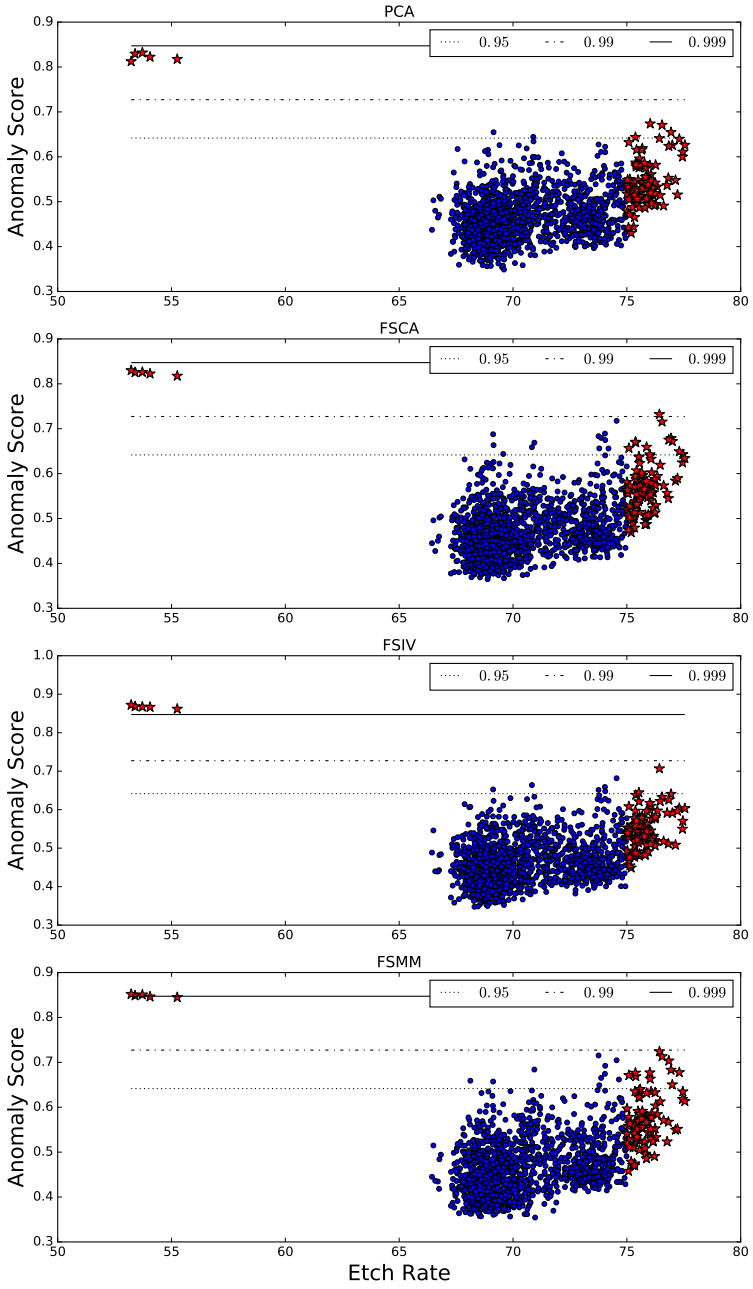

Figure 13: The anomaly score obtained with IF in case study 2 plotted against the Etch Rate value of each wafer. The horizontal lines show the 0.95, 0.99 and 0.999 confidence limits for normal samples. The red stars are the anomalous wafers and the blue dots are the normal ones.

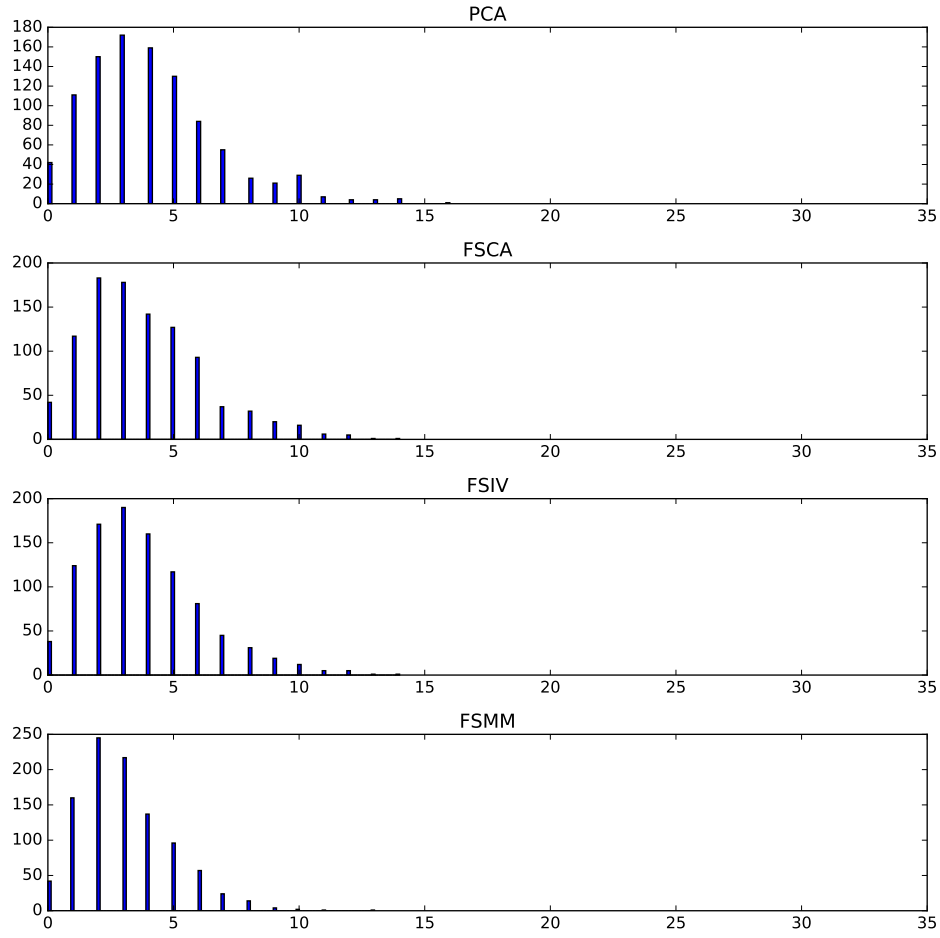

Figure 14: Histograms of the number of splits required to isolate one of the anomalous samples in case study 2 when using IF and the different dimensionality reduction algorithms considered.

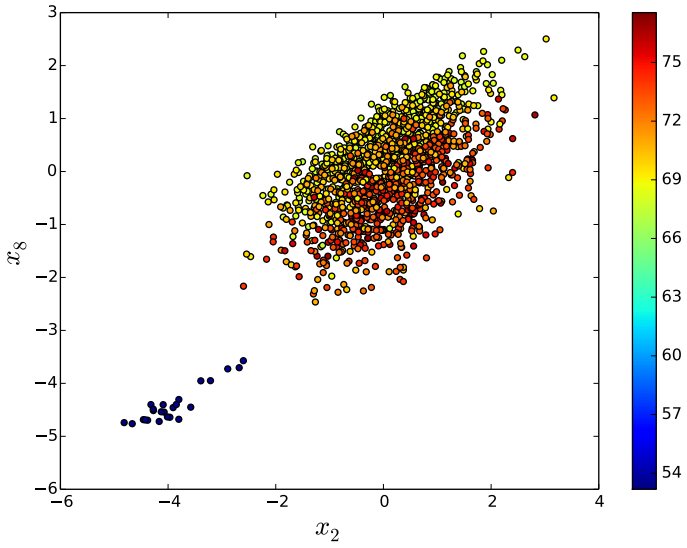

Figure 15: The case study 2 data plotted in the subspace defined by the two variables that were used as a split point in a tree that was able to isolate the anomaly using only 2 splits. The samples are colour coded according to their Etch Rate value.

7. Conclusion

This paper considers the dimensionality and variable correlation problems related to the use of OES data for interpretable anomaly detection in semiconductor manufacturing. Dimensionality reduction tailored to anomaly detection together with IF for anomoly score generation are proposed as an anomaly detection methodology. In particular, FSIV and FSMM are proposed as new feature selection methods that take account of isolated variables in highly correlated high dimension datasets. Both yield better anomaly detection performance than PCA. They also have the advantage over PCA of providing diagnostics that are easier to interpret. In addition, an anomaly diagnosis system based on isolation forest is also proposed that allows individual contributing variables to be identified. The operation and effectiveness of the proposed methods has been illustrated using a simulated example and industrial case studies.

8. Acknowledgment

The first author would like to thank Maynooth University for the financial support provided for the research.

References

Abe, H., Yoneda, M., Fujiwara, N., 2008. Developments of plasma etching technology for fabricating semiconductor devices. Japanese Journal of Applied Physics 47 (3R), 1435.

Bradley, A. P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30 (7), 1145–1159.

Breiman, L., 2001. Random forests. Machine learning, 5–32.

Chen, R., Huang, H., Spanos, C., Gatto, M., 1996. Plasma etch modeling using optical emission spectroscopy. Journal of Vacuum Science & Technology A 14 (3), 1901–1906.

Coburn, J., Winters, H. F., 1979. Ion-and electron-assisted gas-surface chemistry an important effect in plasma etching. Journal of Applied physics 50 (5), 3189–3196.

Flynn, B., McLoone, S., 2011. Max separation clustering for feature extraction from optical emission spectroscopy data. Semiconductor Manufacturing, IEEE Transactions on 24 (4), 480–488.

He, Q. P., Wang, J., Nov. 2007. Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. IEEE Transactions on Semiconductor Manufacturing 20 (4), 345–354.

Jolliffe, I., 2002. Principal component analysis. Wiley Online Library.

Kriegel, H.-P., Zimek, A., et al., 2008. Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 444–452.

Liu, F. T., Ting, K. M., Zhou, Z.-H., 2008. Isolation forest. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, pp. 413–422.

Mahadevan, S., Shah, S. L., Dec. 2009. Fault detection and diagnosis in process data using one-class support vector machines. Journal of Process Control 19 (10), 1627–1639.

Mitra, P., Murthy, C., Pal, S. K., 2002. Unsupervised feature selection using feature similarity. IEEE transactions on pattern analysis and machine intelligence 24 (3), 301–312.

Murthy, S. K., 1998. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. (A2) Data Mining and Knowledge Discovery 2, 345–389.

Prakash, P., Johnston, A., Honari, B., McLoone, S., 2012. Optimal wafer site selection using forward selection component analysis. In: Advanced Semiconductor Manufacturing Conference (ASMC), 2012 23rd Annual SEMI. IEEE, pp. 91–96.

Puggini, L., Doyle, J., McLoone, S., 2014. Towards multi-sensor spectral alignment through post measurement calibration correction. In: 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014). IEEE.

Puggini, L., Doyle, J., McLoone, S., 2016. Fault detection using random forest similarity distance. IFAC-Safe Process 49 (5), 132–137.

Puggini, L., McLoone, S., 2015. Extreme learning machines for virtual metrology and etch rate prediction. In: Signals and Systems Conference (ISSC), 2015 26th Irish. IEEE, pp. 1–6.

Puggini, L., McLoone, S., 2016. Feature selection for anomaly detection using optical emission spectroscopy. IFAC-Safe Process 49 (5), 132–137.

Puggini, L., Mcloone, S., 2017. Forward selection component analysis: Algorithms and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ren, L., Lv, W., May 2014. Fault Detection via Sparse Representation for Semiconductor Manufacturing Processes. IEEE Transactions on Semiconductor Manufacturing 27 (2), 252–259.

Rodgers, J. L., Nicewander, W. A., Toothaker, L., 1984. Linearly independent, orthogonal, and uncorrelated variables. The American Statistician 38 (2), 133–134.

Schaller, R. R., 1997. Moore's law: past, present and future. IEEE Spectrum 34 (6), 52–59.

SIA, 2016. Semiconductor industry association: Global semiconductor sales, online report.

Verdier, G., Ferreira, A., Feb. 2011. Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing. IEEE Transactions on Semiconductor Manufacturing 24 (1), 59–68.

Wilkinson, J. H., 1965. The algebraic eigenvalue problem. Vol. 87. Clarendon Press Oxford.

Yinug, F., 2015. U.s. semiconductor industry employment, jobs issue paper, january 2015.

Yue, H. H., Qin, S. J., Markle, R. J., Nauert, C., Gatto, M., 2000. Fault detection of plasma etchers using optical emission spectra. IEEE Transactions on Semiconductor Manufacturing 13 (3), 374–385.

Zeng, D., Spanos, C. J., 2009. Virtual metrology modeling for plasma etch operations. IEEE Transactions on Semiconductor Manufacturing 22 (4), 419–431.